

Evaluating Models

Metrics for model performance

```
1 augment(tree_fit, new_data = frog_test) %>%  
2   metrics(latency, .pred)  
3 #> # A tibble: 3 × 3  
4 #>   .metric .estimator .estimate  
5 #>   <chr>   <chr>      <dbl>  
6 #> 1 rmse    standard    59.2  
7 #> 2 rsq     standard     0.380  
8 #> 3 mae     standard    40.2
```

- RMSE: difference between the predicted and observed values ↓
- R^2 : squared correlation between the predicted and observed values ↑
- MAE: similar to RMSE, but mean absolute error ↓

Metrics for model performance

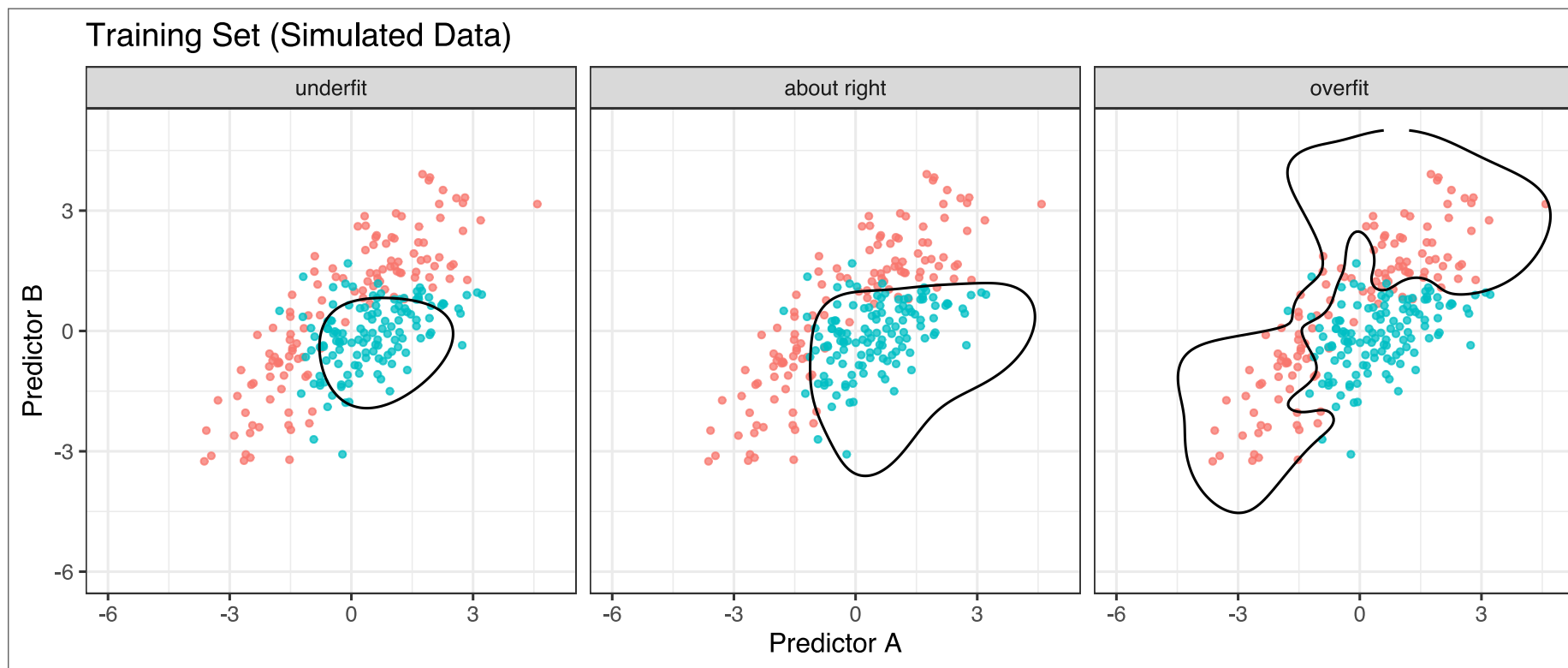
```
1 augment(tree_fit, new_data = frog_test) %>%  
2   rmse(latency, .pred)  
3 #> # A tibble: 1 × 3  
4 #>   .metric .estimator .estimate  
5 #>   <chr>   <chr>       <dbl>  
6 #> 1 rmse    standard     59.2
```

Metrics for model performance

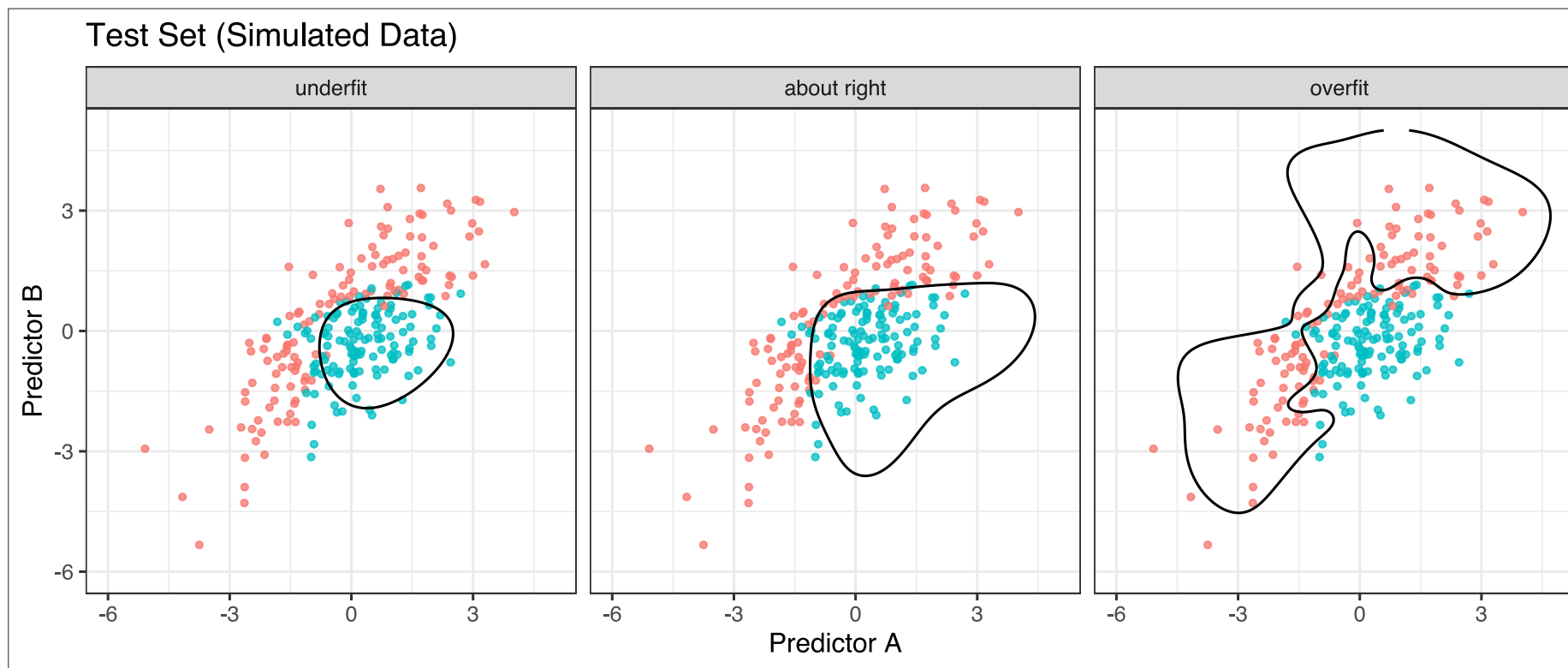
```
1 augment(tree_fit, new_data = frog_test) %>%  
2   group_by(reflex) %>%  
3   rmse(latency, .pred)  
4 #> # A tibble: 3 × 4  
5 #>   reflex .metric .estimator .estimate  
6 #>   <fct>   <chr>   <chr>         <dbl>  
7 #> 1 low    rmse     standard      94.3  
8 #> 2 mid    rmse     standard     101.  
9 #> 3 full   rmse     standard      51.2
```

DANGERS OF OVERFITTING

Dangers of overfitting ⚠



Dangers of overfitting ⚠



We call this “resubstitution” or “repredicting the training set”

What if we want to compare more models?

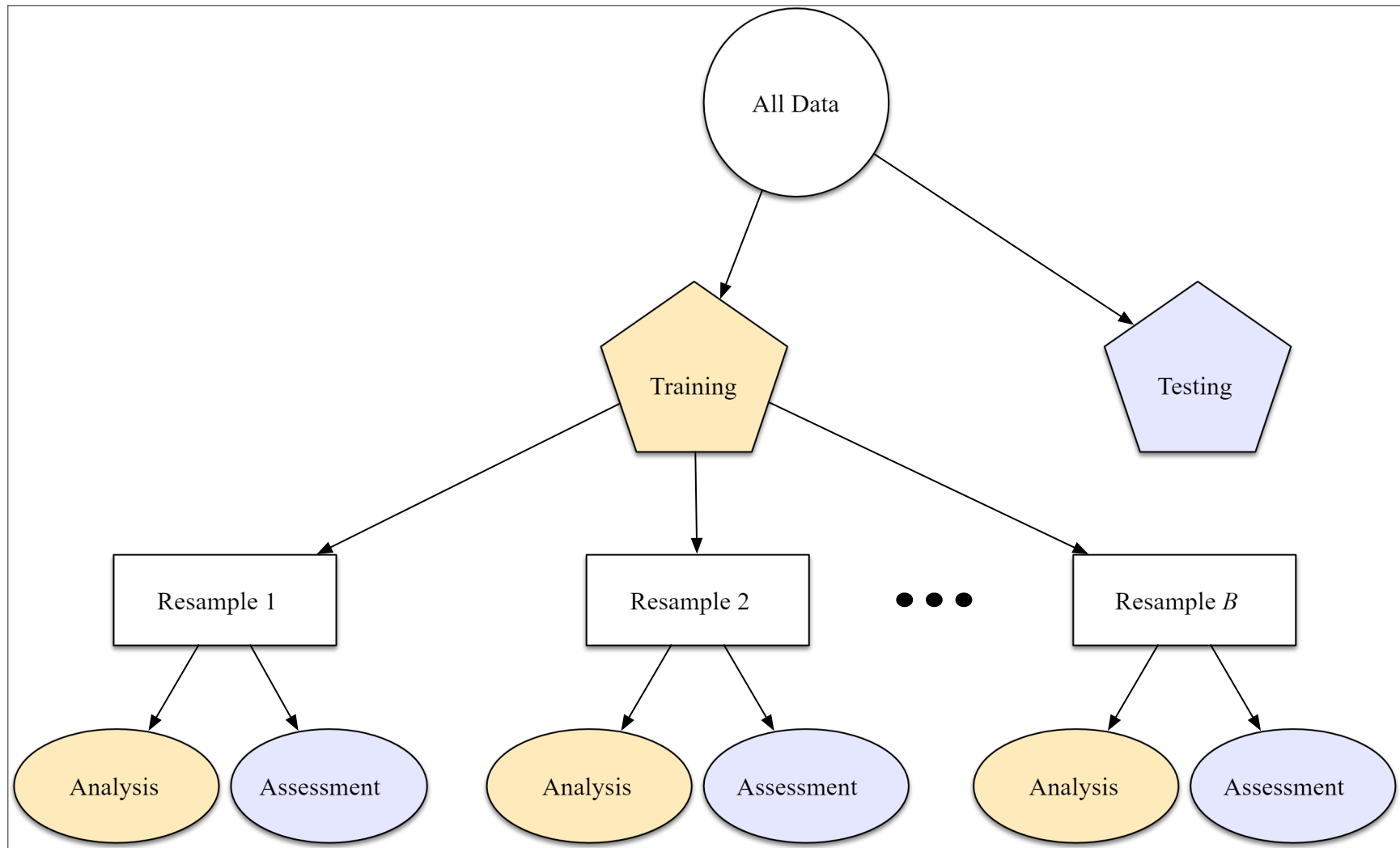
And/or more model configurations?

And we want to understand if these are important differences?

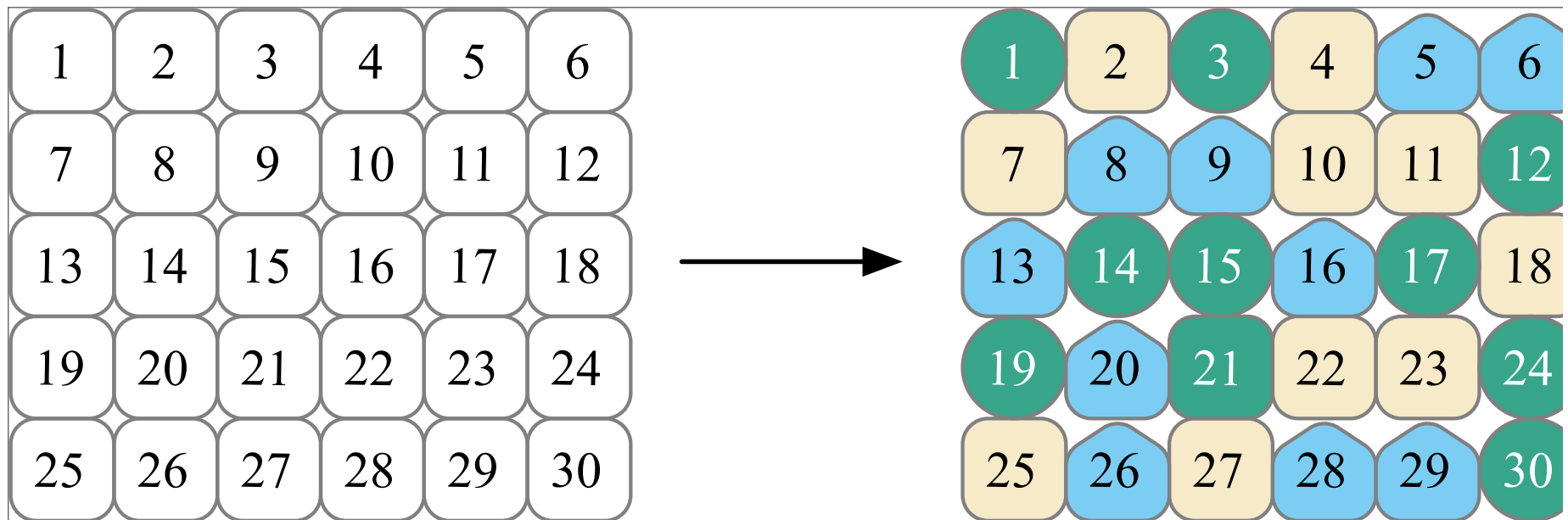
The testing data are precious 

How can we use the *training* data to compare and evaluate different models? 🤔

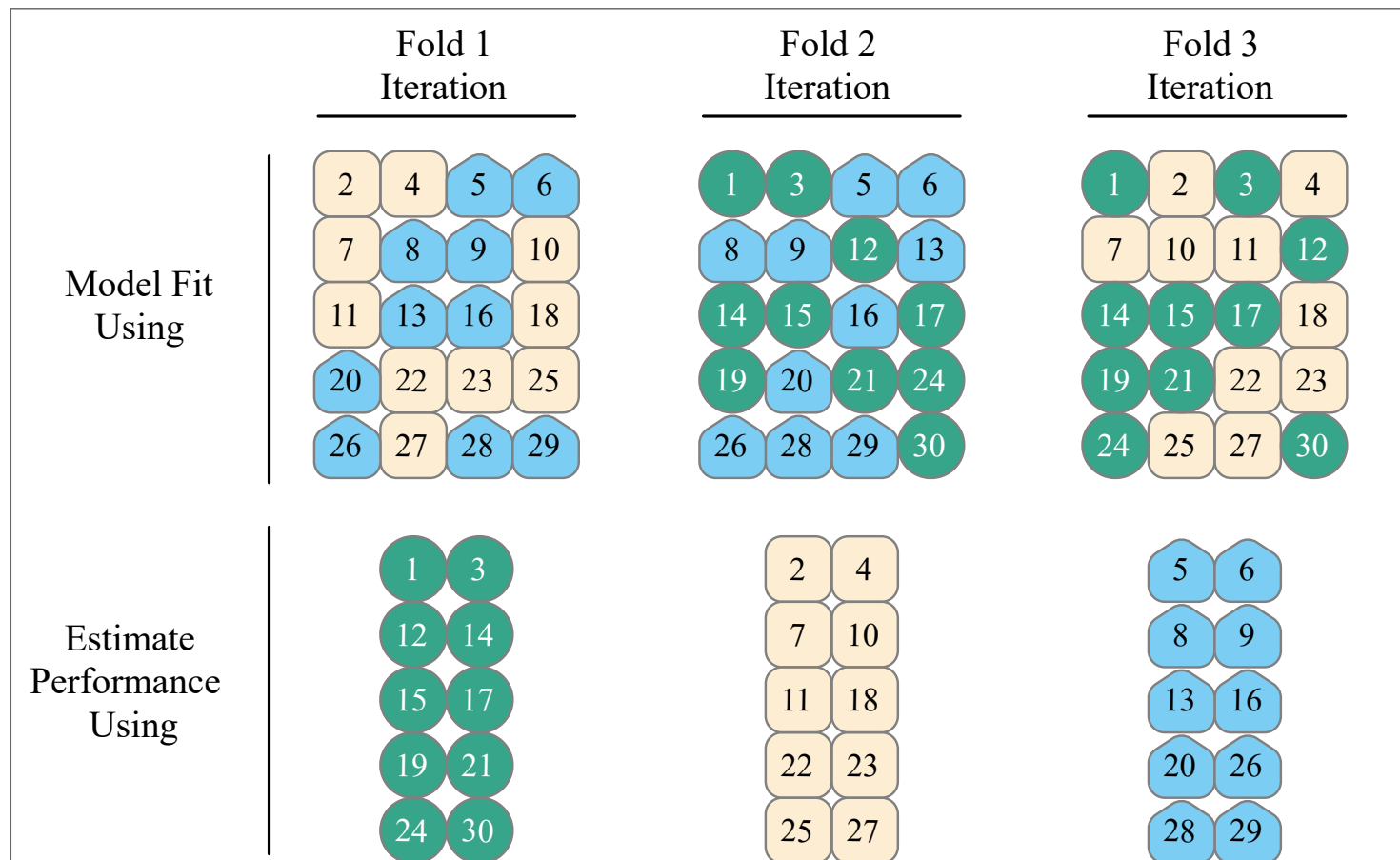
Resampling



Cross-validation



Cross-validation



Alternate resampling schemes

Bootstrapping

	Bootstrap Iteration 1	Bootstrap Iteration 2	Bootstrap Iteration 3
Model Fit Using	<div><div>1</div><div>1</div><div>4</div><div>7</div><div>8</div><div>8</div></div> <div><div>10</div><div>13</div><div>13</div><div>13</div><div>14</div><div>15</div></div> <div><div>16</div><div>16</div><div>16</div><div>17</div><div>19</div><div>19</div></div> <div><div>21</div><div>22</div><div>23</div><div>23</div><div>24</div><div>23</div></div> <div><div>25</div><div>25</div><div>25</div><div>27</div><div>28</div><div>29</div></div>	<div><div>2</div><div>2</div><div>3</div><div>3</div><div>3</div><div>4</div></div> <div><div>4</div><div>4</div><div>6</div><div>6</div><div>7</div><div>10</div></div> <div><div>11</div><div>12</div><div>12</div><div>14</div><div>14</div><div>15</div></div> <div><div>17</div><div>17</div><div>18</div><div>21</div><div>22</div><div>22</div></div> <div><div>23</div><div>23</div><div>28</div><div>27</div><div>28</div><div>30</div></div>	<div><div>2</div><div>2</div><div>3</div><div>3</div><div>4</div><div>5</div></div> <div><div>5</div><div>5</div><div>6</div><div>7</div><div>10</div><div>11</div></div> <div><div>12</div><div>15</div><div>16</div><div>18</div><div>18</div><div>19</div></div> <div><div>19</div><div>20</div><div>20</div><div>20</div><div>21</div><div>21</div></div> <div><div>21</div><div>21</div><div>22</div><div>22</div><div>29</div><div>30</div></div>
Estimate Performance Using	<div><div>2</div><div>3</div><div>5</div><div>6</div><div>9</div><div>11</div></div> <div><div>12</div><div>18</div><div>20</div><div>24</div><div>26</div><div>28</div></div> <div><div>30</div></div>	<div><div>1</div><div>5</div><div>8</div><div>9</div><div>13</div><div>16</div></div> <div><div>19</div><div>20</div><div>24</div><div>26</div><div>29</div></div>	<div><div>1</div><div>8</div><div>9</div><div>13</div><div>14</div><div>17</div></div> <div><div>23</div><div>24</div><div>25</div><div>26</div><div>27</div><div>28</div></div>

Error

×